2018

Structural Principles Analysis of Host-Pathogen Protein-Protein Interactions: A Structural Bioinformatics Survey

Huaming Chen
University of Wollongong, hc007@uowmail.edu.au

William Guo
Central Queensland University

Jun Shen
University of Wollongong, jshen@uow.edu.au

Lei Wang
University of Wollongong, leiw@uow.edu.au

Jiangning Song
Monash University, jiangning.song@monash.edu

Structural Principles Analysis of Host-Pathogen Protein-Protein Interactions: A Structural Bioinformatics Survey

Publication Details

Chen, H., Guo, W., Shen, J., Wang, L. & Song, J. (2018). Structural Principles Analysis of Host-Pathogen Protein-Protein Interactions: A Structural Bioinformatics Survey. IEEE Access, 6 11760-11771.

Structural Principles Analysis of Host-Pathogen Protein-Protein Interactions: A Structural Bioinformatics Survey

Huaming Chen, *Student Member, IEEE,* William Guo, *Member, IEEE,* Jun Shen, *Senior Member, IEEE,* Lei Wang, *Senior Member, IEEE,* and Jiangning Song

Abstract—Computational-intelligence methods in bioinformatics and systems biology show promising potential for leveraging abundant, large-scale molecular data. These methods can facilitate analysis and prediction of the principles of biological systems through construction of statistical and visualised models. Specifically, structural data from exogenous and endogenous protein-protein interactions are of vital significance in this context, encompassing primarily three-dimensional (3D) structural information for a cohort of macromolecules underpinning the biological system. In this study, we surveyed the main methodologies and algorithms for the reconstruction and modelling of the structural-interaction networks (SINs) of host-pathogen protein-protein interactions (HPPPIs), regarding how the protein domains interact with each other to constitute a SIN. Surveying the pattern and organisation of the SIN delivers a state-of-the-art view of HPPPIs and illustrates prospective future research directions. In addition to the binary PPI network, we distilled the relevant data sources into several branching research areas and further expanded the discussions into computational-intelligence methods according to the algorithms applied, including machine learning statistical models, to shed light on effective method design. In particular, atomic resolution level investigations can reveal novel insights into the underlying principles of the organisation and complexity of HPPPIs networks. Combining data analytics and machine-learning technologies, we anticipate that our systematic overview will serve as a useful guide for interested researchers to carry out related studies on this exciting and challenging research topic in system biology.

Index Terms—host-pathogen interactions; structural-interaction network; bioinformatics; machine learning; data analytics

1 Introduction

In this paper, we describe how the computational-intelligence methods can help solve key problems and the dominant mechanisms involved in proteomics research. Considering proteomics represent the large-scale study of proteins, proteomics relies upon the investigation of several aspects, including when, where, and how proteins function, and how proteins interact with each other. Recently, an abundance of experimental data has accumulated, propelling hypothesis-driven biomedical research into the big-data era.

Given the continuous growth and availability of large-scale multi-omics data, both the protein-protein interaction (PPI) networks and structural analyses involving proteomics remain hot topics. Exploration of proteomics data sources, such as those from the European Bioinformatics Institute [2], [3], [4], promotes research in transforming biomedical research at system-level, mechanistic studies aimed at a comprehensive and holistic understanding of biological systems [5]. Although challenges, such as specialised domain knowledge and data issues, might hinder proteomics researches, this data-driven work to obtain extensive information about systems from large amounts of raw data is currently popular in both academia and industry [6].

Systems biology [7] represents the comprehensive study of presenting a holistic view and analysis of biological processes. Specifically, systems biology aims to understand and further predict the behaviour of biological systems [8] and includes studies on functional genomics and proteomics. There are several studies focusing on genomics data, mostly from The Cancer Genome Atlas (TCGA) [9], given that a nearly complete map for human and other species had been provided along with the development of genome-sequencing projects [8]. These studies provided insights into gene-related networks and a fuller understanding of how a set of molecules interacts with each other [10]. Three-dimensional (3D) structures of these molecules are the most critical for deriving relationships.

Our study was focused on proteomics, and specifically on HPPPIs. Considering the prevalence of protein interactions between species, most early studies were performed within the same species due to the limited availability of proteomics data at that time [11], [12]. Several recent studies demonstrated improvements in PPI between different species, which were referred as "interspecies PPI", and that offered important information for further analysis of infectious mechanisms [8], [13]. However, beyond the interaction between these PPIs, their structural informa-

- *H.Chen, J.Shen and L.Wang is with the School of Computing and Information Technology, Faculty of Engineering and Information Science, University of Wollong, NSW, 2522, Australia*
  E-mail: hc007@uowmail.edu.au; {jshen, leiw}@uow.edu.au
- *W.Guo is with the School of Engineering and Technology in Central Queensland University and J.Song is with the Department of Biochemistry and Molecular Biology in Monash University.*
  E-mail: w.guo@cqu.edu.au; jiangning.song@monash.edu
- *This paper is an extension paper from its conference version [1].*

Manuscript received ; revised .

tion is vital to their discovery. We anticipate that study of the identified data collected via open databases [14] would present a comprehensive survey towards structural principles concerning the PPI identified between the host and pathogen. These HPPPIs are experimentally verified and manually recorded in systems and include information regarding infection pathways in their interaction networks and are able to reveal much more information regarding the infectious mechanisms between hosts and pathogens. We first investigated a previous HPPPI study [14] and expanded our work based on the preliminary sequence information [13], [15] to exploit the online available and experimentally verified HPPPI data. However, these studies simply focused on binary protein interactions prediction.

In addition to these studies, we expect to leverage the structural information of the HPPPI data for building structural-interaction networks (SINs) with respect to simply classifying pairs of proteins as interacting or not. The structural information of the HPPPIs represents various protein properties, from which systems biology might extract a highly convincing network-analysis result and introduce trustworthy statistics in cooperation with the corresponding structural information and domain data, as well as the atomic resolution-level networks.

Therefore, the structural-principle analysis of HPPPI networks is discussed and surveyed in the following sections, which covers most branches closely associate with the protein structural information. This analysis was achieved by SIN, an atomic-resolution PPI network [16]. Protein structural information is another experimentally determined set of 3D data previously described. It mainly contains several protein properties, including domain information, family annotation, secondary/tertiary structure.

Because there are few 3D-specific studies offering an atomic view of HPPPIs, we provide an overview of progress made by biologists in relation to bioinformatics, including 3D structural databases and analysis based on the structural information. Our efforts will help readers navigate gaps between biological analysis and computational modelling. This includes:

- Protein secondary/tertiary structure prediction
- Domain-domain interaction prediction

These provide the basics for reconstruction of a SIN.

The remainder of the paper is organised as follows. We firstly present the preliminary concepts in Section 2, including the sequence information and the representation algorithms, structural information and domain-domain interaction. Section 3 lists the public repositories and databases. In Section 4, a variety of machine-learning algorithms developed and applied for protein-structure analysis and domain prediction are discussed, and a detailed process to layer curated 3D structural models on top of the binary interaction network is described in Section 5. Section 5 also provides a linkage between model knowledge and analysis. The challenges to building a structural interaction model are discussed in Section 6, and we conclude the review in Section 7.

2 Preliminary Concepts

The two main predictive tasks associated with proteomics related to computational biology are the protein structure and the domain-domain interaction. Both sets of data are usually difficult for bioinformatics researchers to obtain; however, building a SIN requires a complete understanding of both protein structure and domain features. In this section, we present the biological meaning for both the structural information and the domain-domain interactions, and also introduce the modelling process necessary for completing the prediction of both tasks.

2.1 Sequence Information

Proteins are comprised of various numbers of amino acids as their basic building blocks. The concatenated string of amino acids forming the folded protein represents its primary sequence information. Typically, there are 20 different proteinogenic amino acids [11], although five additional amino acids exist in the human and pathogen protein sequences [14], including selenocysteine/U, pyrrolysine/O, aspartate or Asparagine/B, glutamate, and glutamine/Z.

Figure 1 shows the 20 different amino acids.

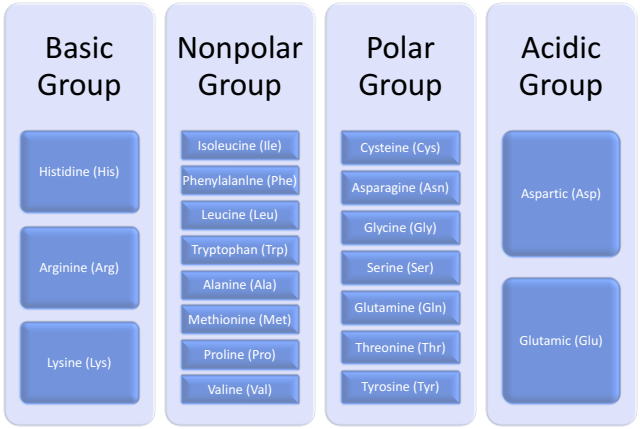

Fig. 1. Amino Acids Groups

As a preprocessing step for inputting sequence data into computational model built for protein classification and regression tasks, transformation of efficient and effective data into the model is necessary. Sequence representation is a vital preprocessing step for efficiently and effectively feeding data to any computational model for protein classification and regression analysis. In TABLE 1, we list several mainstream algorithms concerned with sequence representation, where the protein sequence is denoted as $X = x_1, x_2, ..., x_n$. We define the amino acid number as 20 for this paper.

These different sequence-representation algorithms provide as much information as possible to the computational model in different vector lengths. Because the sequence information is easier to obtain via the high-throughput technology, it is primarily utilised for both protein structure prediction and interaction prediction.

TABLE 1
Protein Sequence Representation Algorithms

Algorithm	Reference	Definition	Prefix	Equation	Feature Dimension
Amino acid composition	[17]	Each feature represents the frequency of the corresponding amino acid type in the protein	aa_i is one of the 20 types of amino acids $aa_1, aa_2, ..., aa_{20}$	$f_i = \frac{counts_{aa_i}}{n}$	$1 * 20$
Conjoint triad method	[11]	Considering the properties of one amino acid and its vicinal amino acids as a pattern f_i, the frequency of f_i represents one feature. The concatenation of these f_i defines a unique feature vector.	For the amino acids that have been catalogued into seven classes, $F = f_1, f_2, \ldots, f_{343}$. $D = d_1, d_2, \ldots, d_{343}$	$d_i = \frac{f_i - min(f_1, f_2, ..., f_{343})}{max(f_1, f_2, ..., f_{343})}$	$1 * 343$
Auto covariance	[12]	Projecting the amino acids with their specific seven kinds of physiochemical properties, auto covariance formalizes the sequence information into a uniform matrix	$P_{i,j}$ is the jth property of the ith amino acid, while the protein has n amino acids. lag is defined as the distance between two amino acids and lg is the maximum value of lag.	$AC(lag, j) = \frac{1}{n-lag} \sum_{i=1}^{n-lag} (P_{i,j} - \frac{1}{n} \sum_{i=1}^{n}) * (P_{i+lag,j} - \frac{1}{n} \sum_{i=1}^{n} P_{i,j})$	$lg * 7$
Local descriptor	[18]	Segmenting a protein sequence into several individual regions, i.e. 10 regions in [18], three descriptors are used to describe each region, including Composition (C), Transition (T) and Distribution (D).	The basis to group amino acids is considered by different biology schemes, i.e. three functional groups (hydrophobic (CVLIMFW), neutral (GASTPHY) and polar (RKEDQN)), seven physiochemical groups.	$i = 1, 2, \ldots, 7$; $c_i = \frac{counts_{C_i}}{n}$; $t_i = \frac{counts_{T_i}}{n-1}$; $d_i = \frac{loc(T_i)}{n}$, $loc(T_i)$ denote the location index of i	$1 * 630$
Position-Specific Scoring Matrix (PSSM)	[19]	The defined matrix, P, is in $n * 20$ dimensions, where $P(i, j)$ indicates the possibility of the jth amino acid appears at i position. PSI-BLAST [20] is one of the most frequently used tools. PSI-BLAST [21] is one of the most frequently used tools.	The protein sequence is divided into 20 blocks while its length is n.	$P_{ij} = \sum_{k=1}^{20} w(i, k) * Y(j, k)$; $F_j = \frac{1}{B_j} \sum_{i=1}^{B_j} P_i^(j)$	$20 * 20$
One-hot sparse vector	[22], [23]	Each amino acid is defined in a one-hot sparse vector. The length, M, of vector is dependent upon the number of the amino acid types, i.e. 25 in [14], 22 in [22] and 21 in [23]	Normally, a balance cut-off value should be defined before preprocessing. *700* is mostly used.	Each row only has one position with value '1'	$[700 * 20]$

2.2 Structural Information

Because protein sequences exhibit various lengths, those with < 50 amino acids are generally referred to as polypeptides and contain only primary level information. For secondary structure, folding forms common structures, such as $\alpha - helices$ and $\beta - sheets$ (from $\beta - strands$). Another structure is referred to as a random coli. Upon folding, a secondary structure subunit transforms into tertiary structure. For some proteins, their structure consist of more than one polypeptide, suggesting multiple tertiary structures. This context information is subsequently referred to as quaternary structure. We illustrate the 3D structure for protective antigen (UniProt ID: 'P13423') in Fig. 2.

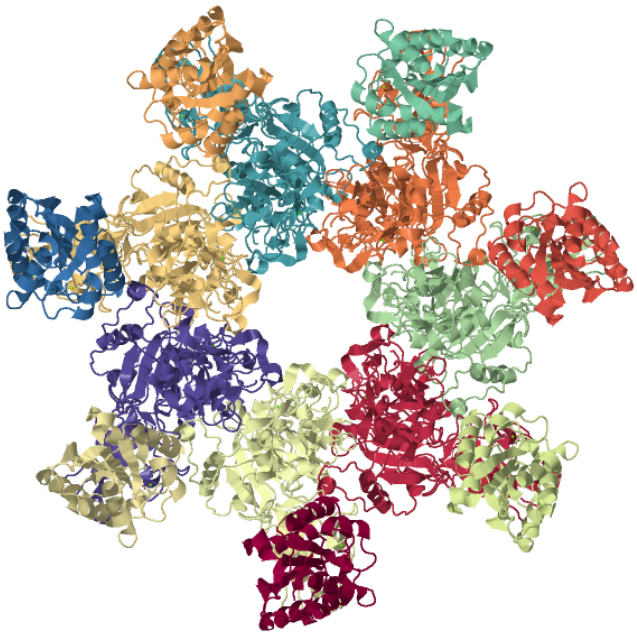

Fig. 2. The 3D structure of the protective antigen (Uniprot ID: 'P13423')

Because the wet lab is the site of protein-structure determination by X-ray crystallography, NMR spectroscopy or cryo-electron microscopy, these methods are extremely time-consuming and expensive. Therefore, an *ab initio* method based on computational modelling is a current focus of academic and industrial research. Only $< 0.5\%$ of all sequenced protein structures have solved structures according to the limitations of biological experiments methods [24].

Study of secondary structure prediction creates a dictionary of protein secondary structure (DSSP), which is better defined and clearer than tertiary structure and quaternary structure. Additionally, secondary structure can be analysed using efficient sequence information from primary structure. The secondary structure is predefined with three types of motifs: α-helix, β-strand and coli, allowing Q3 accuracy [23], [25], [26], [27]. Statistical models and machine-learning methods have extensively improved Q3 predictive accuracy from 65% to 80%. Recently a more challenging problem targeting on eight-category prediction (Q8) defined in DSSP for secondary structure prediction was described. These eight categories describe the secondary structure based on additional elements: 3_{10}-helix, α-helix, π-helix, β-strand, β-bridge, β-turn, bend and loop/irregular [22], [28]. To achieve more accurate results on secondary structure, these methods require not only an efficient model but also sufficient feature representations from the sequence information. The involved models will be introduced in Section 4. The key challenge to predicting secondary structure involves prediction of those proteins having no close homologs and that have not experimentally verified 3D structures.

To achieve sufficient feature representations for secondary structure prediction, most studies introduced the protein-sequence information, amino acid profile information, local and global sequence information [23], [26], [29], [30]. In this study, we first focus on the eight categories for secondary structure prediction

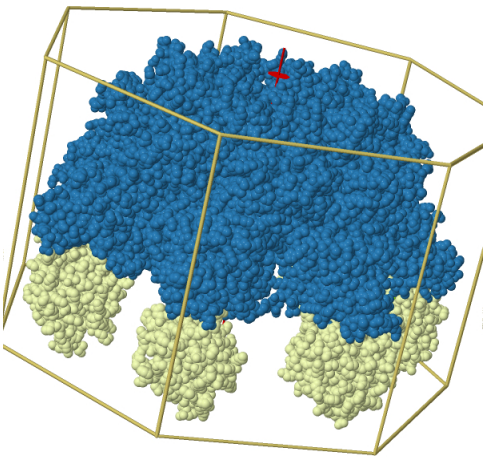

Fig. 3. Tertiary Structure of protein Protective Antigen (Uniprot ID 'P13423')

Fig. 3 provides an example of a tertiary structure of the protective antigen protein (UniProt ID: P13423). Prediction for this level of structure normally involves homology modelling [31], which is also known as comparative modelling, where the main resulting candidate is derived from amino acid sequence alignment by mapping amino acids between different sequences. Introduction of homology modelling method into tertiary structure prediction allows evolutionary results to reveal proteins harboring similar amino acid sequences based on their shared similar tertiary structure to accomplish related biological function [32].

The structure information is a requisite for structural interaction networks, given that they provide atom level information. In Section 3, we will describe related databases available for acquiring such information.

2.3 Domain-Domain Interactions

Given a protein sequence, protein domains are distinctive functional or structural subsegments. Most protein domains build independently stable and folded 3D structures, with which the domains combined into different arrangements to form a unique protein with different functions [33]. Therefore, binary PPI networks can be further considered at the domain level, especially when the interacting protein is large. Although most proteins consist of multiple domains, a

pair of PPIs often involves only one pair of domain-domain interaction focusing on the actual binding site.

Domain-level interactions provide a global view of the binary PPIs network. For HPPPI investigations, this reveals interaction location or pathological interactions and can help facilitate drug-development targeting for infectious diseases. To acquire a comprehensive understanding of how domain interactions are mediated, the primary method involves analysis of individual interactions using experimentally determined 3D structures. However, this information is available for only a small fraction of proteins, indicating the domain-level PPI data not readily accessible.

There are several existing databases, including 3did [34] and iPfam [35], that provide domain-domain interactions by identifying these based on experimentally determined 3D structures. Other databases provide combined interactions, in which data are derived experimentally and the rest is computationally predicted. DOMINE [36] includes both 3D-structure-based and predicted domain-domain interactions and shows the predicted domain-domain interactions at three different levels, namely 'High', 'Middle' and 'Low'. Two primary methods, association [36] and maximum-likelihood estimation [37], are introduced in this domain-domain interaction-prediction task. The essential information utilised in these models includes domain information from protein sequence and binary PPI information.

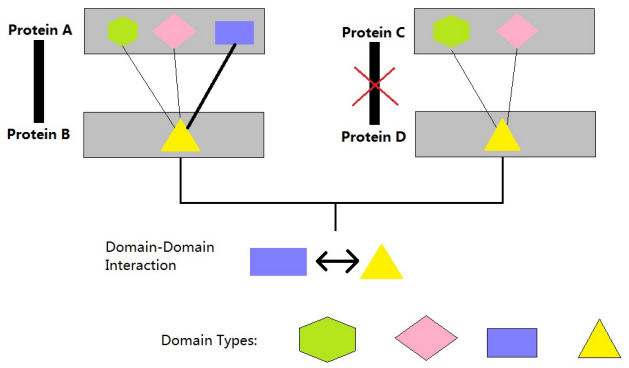

Fig. 4. Domain-domain Interaction

To provide a general understanding of domain-domain interactions associated with binary PPIs, Fig. 4 shows a basic diagram for domain-domain-interaction prediction [38]. 'Protein A' interacts with 'Protein B' while 'Protein C' does not interact with 'Protein D'. Several different domains types are identified using the related databases. Mostly, we choose Protein Data Bank (PDB) [39] as suggested. Later, we will compare differences between these two groups of domain-domain relationships to identify the interacting domains between two different proteins.

3 RELATED DATABASES

Ranging from protein-sequence information to their structure data, several different databases are currently available and well maintained, including host-pathogen PPI databases, structure databases, protein families and domain databases, and also domain-domain-interactions databases.

3.1 Host-Pathogen Interactions Databases

Although several different standardized formats for the host-pathogen PPIs are published by different organizations, these databases contain the most important binary information for HPPPI researches. Some popular repositories are initially built by universities, such as HPRD by Johns Hopkins University and the Institute of Bioinformatics, PATRIC by University of Chicago, PHISTO by Boazii University, VirHostNet by Universit de Lyon. Highly credible positive HPPPI pairs are manually recorded in these systems and updated periodically. The details of several popular databases are listed in TABLE 2.

TABLE 2
Host-Pathogens PPIs Database

HP-PPI Database	Contents
HPRD [40]	A database manually extracted from literature, is built by Johns Hopkins University, includes more than 39,000 interaction pairs.
BIND [41]	It belongs to Biomolecular Object Network Databank, and is maintained by University of Toronto. It provides more than 200,000 interaction pairs.
DIP [42]	It includes several sources, i.e. Yeast Protein Database, Kyoto Encyclopedia of Genes and Genomes.
PATRIC [43]	Continuously updated by University of Chicago, this database is built upon a combination of several public repositories.
PHISTO [44]	Currently it stores over 23,000 interaction pairs and these data are imported from several PPI databases using PSICQUIC tool.

3.2 Structure Databases

The Protein Data Bank (PDB) [39] is the primary database housing structural information for proteins and is managed by the worldwide Protein Data Bank (wwPDB) international collaboration. The PDB contains all experimentally determined protein structures ranging from different resolutions and detection methods.

The PDB is currently updated weekly and has its own file format standard, which is strictly defined to provide protein and nucleic acid structure details. A standard PDB file should contains atomic coordinates, observed sidechain rotamer, secondary structure assignments and atomic connectivity information. Apart from the critical information, abbreviation content about the corresponding literatures is also mandatory in PDB file, which is listed as Header. Several other specific columns include the ID number, date for publication, obsolete status, details about the related experimental methodology, molecular components of the complexes, the source of the complexes, the experimental method used to determine the structure, the authors, modification and revocation records, and related literature, the maximum resolution, and other statistics.

A simple example of the protective antigen protein (UniProt ID: P13423) using PyMOL [45], [46] is shown in Fig. 2. It requires substantial time and effort to acquire an experimentally determined protein structure, and currently, not every protein has its corresponding structural information available. Determination of this information for these proteins is critical for building a SIN.

3.3 Protein Families and Domain Databases

As an important database of protein domains and families, Pfam provides a complete map for protein domains and families [47], [48]. It is regularly updated, with the latest version being Pfam 31.0 released in March 2017 for instance and containing >16,712 protein families.

Although amino acids are the elements comprising a protein sequence, functions occur in multi-sequential regions which are called domains. Identifying these domains provides details and insights regarding the functional mechanism of the protein.

Structural information allows bond information detailing interactions between proteins, which is more concrete than binary HPPPI network provided in HPPPI databases. Therefore, iPfam is used in SIN studies to identify domain-domain interactions between proteins [35]. iPfam was developed by Howard Hughes Medical Institute, and currently harbors $>$ 9,500 domain-domain interactions.

iPfam is based on two continuously updating databases, PDB and Pfam, both of which are well established for their 3D structure and domain-information purposes. Most of the structural information in the PDB also contains multiple domains. The 3did is another domain-domain interaction databases for 3D-interacting domains between proteins, and is a collection of protein interactions from which high-resolution 3D structures are known [34], [49].

By using iPfam and 3did to achieve domain-level resolution of HPPPIs, SIN considers proteins in their precise spatial relationships by layering domain-domain interactions on top of the conventional PPI networks. As protein-sequence information accumulates at a staggering rate, these data depict its characteristics with high volume, high velocity, high variety, high value and high veracity (5V). This, along with big-data analytics, including machine-learning technologies, allows addressing structural and domain-domain-interaction prediction problems. In the following section, we introduce the related computational models or methods for SIN construction, including machine-learning methodologies.

4 Computational Models

SIN is designed to layer high-confidence 3D models on top of PPIs. Before layering the structural information on the binary HPPPI network, the structural information of corresponding proteins is requisite. However, only a few proteins have experimentally determined structure, specifically with high-resolution scale. Therefore, herein we present related studies outlining structure prediction and domain-domain-interactions prediction. We review this section as an important step in jointly studying protein structural information while supplementing the structural interaction network.

4.1 Bayesian Statistics

The earliest studies on protein secondary structure prediction mainly focused on the use of Bayesian statistics [50], [51], [52]. Basically, Bayesian statistics describes this problem as follows:

$$I(S;R) = log[\frac{P(S|R)}{P(S)}] \quad (1)$$

where $P(S|R)$ is the conditional probability for observing a conformation S, when a residue (amino acid) R is present, and $P(S)$ is the probability of observing S. According to the conditional probabilities definition, $P(S|R) = P(S,R)/P(R)$. $P(S,R)$ is the joint probability of S and R. Through the use of Eq. (1), an estimation of $I(S;R)$ from a database of known protein sequences and corresponding secondary structures can be achieved.

Specifically, a previous study [51] showed that the the Garnier-Osguthorpe-Robson (GOR) method based on information theory used a 17-amino-acid sequence window to extract properties from protein sequences. The GOR method presented the observed frequencies of singletons, then in pairs of residues on a local sequence of 17 residues to build the Bayesian model, followed by estimation of the probabilities for the Q3 structures. This method increased the accuracy from 55% to 64.4%. Later, in [52], combined with information theory, GOR V algorithm projects the known twenty amino acids types for each specific secondary structure to achieve a Q3 accuracy of 73.5%.

4.2 Support Vector Machine (SVM)

Using SVMs to predict protein secondary structure was firstly introduced in 2001 [53], with the first SVM proposed in 1995 [54]. It is not the first machine learning approach used for protein secondary structure prediction, yet by then, it achieved the best performance overall on Q3 task.

Similar to earlier researches using neural network based methods [29], the encoding scheme for the input layer is called a local-coding scheme and denotes every amino acid with a 21-dimensional orthogonal binary vector as follows:

$$(1,0,,0) or (0,1,,0), etc$$

In the output layer, the Q3 task was first considered as a binary classifier later combined into a tertiary classifier.

A previous study [53] considered the SVM as a superior model based on its ability to effectively avoid overfitting and to handle large feature spaces. In details, the authors [53] selected the radial basis function as the kernel function to train the SVM, resulting in a Q3 task of 73.5%.

4.3 Random Forests

Apart from predicting secondary structure, domain-domain interaction is also critical to the SIN. The random forest model was introduced to build multi-classifiers to determine a decision for a dataset with 1050-dimensional features [55]. Additionally, another study [56] showed an ensemble model of random forests and SVMs were able to predict the domain-interacting sites.

Derived from decision trees model, random forest leverages the power of randomisation to increase model performance [57], [58]. It is able to deal with imbalanced data

problems via the voting mechanism while its random feature selection benefits the model in case of high-dimensional data.

4.4 Artificial Neural Networks

To the best of our knowledge, artificial neural networks were first introduced in protein secondary structure prediction using a fully connected three-layer network in [29], with a learning algorithm involving back propagation. Later, the authors of [59] used a two-tier architecture to deploy neural networks for prediction; however, the improvement in Q3 accuracy has since stalled.

Recently, Q8 accuracy has been the focus of academia and industry, aiming to apply deep learning techniques to improve performance. In [60], probabilistic graphical models, which combine conditional neural fields (CNFs) with neural network, were deployed to improve Q8 accuracy. The features are extracted from position-specific score matrix (PSSM) and the physico-chemical properties of the amino acids. Both the complex relationship between sequence and secondary structure information, and the interdependency relationship among secondary structure types of adjacent amino acids were studied using the CNFs model [60].

Generative stochastic networks (GSNs) were utilised to learn a generative model of data distribution without explicitly specifying a probabilistic graphical model [22], . Specifically, this supervised extension of GSNs is deployed via learning a Markov chain to sample from a conditional distribution for training on a protein structure prediction task. This model was presented with deep learning techniques to tackle Q8 problem for protein secondary structure prediction. The empirical design for the data preprocessing step involved choosing 700 lengths as the cut-off threshold to balance the efficiency and coverage of protein sequence. The main features extracted included the evolutionary information (PSSM feature) and the sequence information (one-hot binary vector feature). The model achieved 66.4% accuracy on Q8 problem.

The most recent result on Q8 accuracy task was reported in [23], which proposed a deep convolutional and recurrent neural network. The feature encoding the protein sequence remained partially similar to the local-coding scheme. In this network model, a feature embedding layer was deployed to map sequence information and profile feature (by PSI-BLAST) to a denser matrix. Multiple convolutional neural network layers and stacked bidirectional relational neural network layers were included to learn both local context information and global context information from the denser matrix. Fully connected and softmax layers were layered on the top of the model to build the classifier for the prediction task.

Considering the different properties of protein structure, an iterative use of predicted features, including the backbone angles and dihedrals based on C_α atoms, improves secondary structure prediction accuracy [61]. Stacked sparse auto-encoders with three hidden layers were introduced. The hidden layers were all with 150 neuron nodes. The method achieved an accuracy 80.8% in secondary structure prediction in the recent CASP targets[1] [61].

1. http://predictioncenter.org/casp11/index.cgi

Various models have been discussed in this section; however, our goal is to stack these different data types atop the binary HPPPI network to achieve structural principles analysis. In the following section, we will discuss the structural interaction network.

5 STRUCTURAL INTERACTION NETWORK

Since principles analysis of protein interactions between host and pathogens still remains poorly understood, an ensemble network of binary HPPPI networks and structural information would provide an efficient option for mining this knowledge using a systems biology approach.

A previous study used 3,949 genes, 62,663 mutations and 3,453 associated disorders for analysis using a 3D structurally resolved human interactome network [62]. By integrating data from iPfam, 3did and the Human Gene Mutation Database (HGMD) [63], a high-quality binary PPIs network with the atomic-resolution interfaces was successfully built [62], providing key insights to in-frame mutations, locations, and disease specificity for different mutations in the same gene, which had not been possible to be acquired on a low-resolution network. The original interaction network obtained from literature-curated databases [62] contained 82,823 pairs; however, after filtering out the proteins without experimentally determined structures, only 4,222 structurally resolved interactions between 2,816 proteins remained. To build a structural interaction network still requires more efforts on experimental determination of a structure or computational prediction, because only a tiny fraction of these binary PPIs can be analysed with their corresponding structure information.

Our previous study [14] collected all the experimental protein interaction data from the published databases, among which we chose the databases being manually checked and uploaded. TABLE 3 shows the five bacterial species with HPPPI statistics. The HPPPI network is further illustrated for *Clostridium botulinum* in Fig. 5[2] [44].

TABLE 3
Statistic of HP-PPI Data Set

Bacteria Species	Positive Pairs Number
Bacillus anthracis	3138
Clostridium difficile	53
Escherichia coli	104
Francisella tularensis	1339
Yersinia pestis	4118

Fig. 5 shows six primary human proteins interacting with nine *Clostridium botulinum* proteins, resulting in 44 HPPPI connections derived from the PHISTO database. These interactions are considered as exogenous interactions. To further analyse interactions from the PPI network, we embedded this information with structural information. There are two classes of protein-protein interaction in physical interactions: interactions mediated by two domains and that between short motifs and domains.

2. http://www.phisto.org/index.xhtml

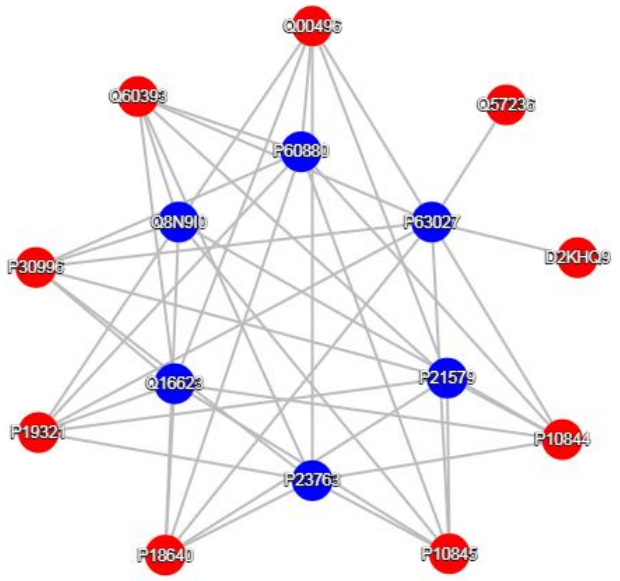

Fig. 5. Binary PPI Network of Clostridium botulinum

We observed that several possible structural principles analyses were obtained within the human-virus protein-protein interaction network [16]. The SIN approach in human-virus PPIs network reveals atomic resolution, mechanistic patterns, and allows systematic comparison with human endogenous interactions.

Figure 6 shows an example detailing how to layer the structure and domain-domain interaction information on top of the binary PPIs network [16], [64].

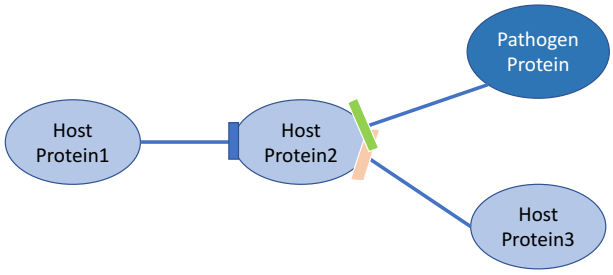

Fig. 6. Structure Interaction Network [64]

Figure 6 reveals the overlapping interfaces between the "Pathogen Protein-Host Protein2" and the "Host Protein3-Host Protein2", which determine the interaction. This type of information could not be observed in the binary PPI network. Further analysis revealed that "Pathogen protein" is mimicking the action of "Host Protein3". Layering the 3D structural information to illustrate the details of the protein interaction allows derivation of two different classes of protein interactions (Fig. 7 and Fig. 8) [65]. The results are generated by PyMOL [46].

The illustration examples present the non-overlapping protein-protein interactions by 3D structures *1F5Q-1BUH*, and overlapping protein-protein interaction by *4MI8-2P1L* [65]. Here, *1F5Q*, *1BUH*, *4MI8* and *2P1L* are their PDB id.

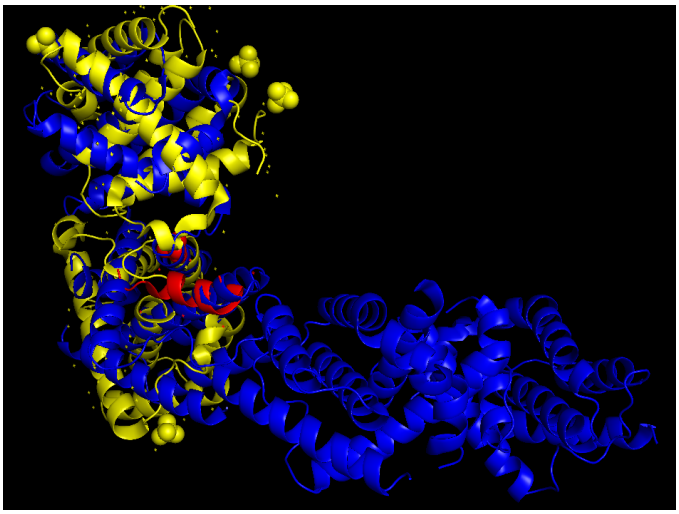

Fig. 7. The Overlapping Structure Interaction: The red string is the human protein Beclin-1, which is annotated with *5EFM* as its PDB id. The compound (in yellow), which is interacted by human protein "Beclin-1" and Gamma Herpesvirus protein "v-Bcl2", is associated with the compound (in blue) by human protein "Beclin-1" and human protein "BCL-XL". The 3D structure of yellow compound can be fetched by PDB id *4MI8* while the blue is *2P1L* [65].

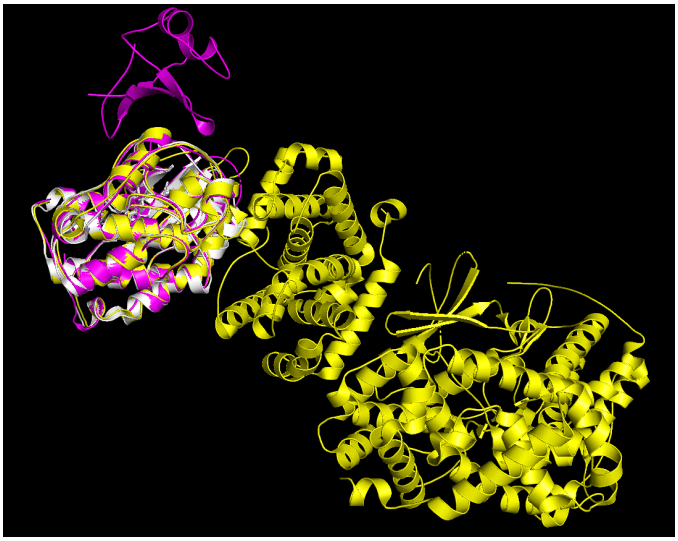

Fig. 8. The Non-overlapping Structure Interaction: The interaction is linked by the human protein "CDK2". The PDB id is *5MHQ*. The yellow compound is the interaction between Gama Herpervirus "Cyclin" and human protein "CDK2". The purple compound is by human protein "CKS1" and "CDK2" [65].

The host-pathogen PPI networks provide specific pathogen protein functions and the global analyses on this network help revealing critical proteins in the networks [64]. Although Fig. 6 provides essential mappings via the overlapping interfaces, annotating the experimental HPPPI networks with 3D structural information will provide further information, because the PPIs can be combined between two globular domains and also between one short linear motif (a short functional segment considered on secondary

structure) and globular domains. Superimposing structures of the HPPPI can help to visually reveal the details.

Several methods to assemble structural information with binary HP-PPI network include:

- Using only the experimentally determined structural information. Both proteins in the HPPPI network could be mapped along with the determined structural information;
- Using both the experimentally determined and computationally predicted structural information. One of the proteins in the HPPPI could not be mapped with its determined structural information;
- Using only the computationally inferred structural information. Both proteins in the HPPPI could not be mapped with its determined structural information. The homology modelling method is widely used for searching for homologous proteins with having determined structure according to the BLAST E-value.

Computationally predicted structural information mainly comes from homology modelling, which is widely used in bioinformatics, provided that protein structure and function are primarily determined according to their sequence information [16].

Typically, for host-pathogen protein-protein interactions, we hypothesised that imitating the binding activities between proteins would allow insight into primary mechanism associated with infections. Given a SIN, there are several types of statistics data that may help us propose and support this hypothesis. As a specific example between virus and host-PPI networks, a previous study [16] analysed the exogenous and endogenous interactions in the human-virus SIN model.

Meanwhile, the overlapping ratio of protein interactions involved in exogenous interface to those involved in endogenous interface indicates potential infectious targets, although the mapping of endogenous interfaces is not guaranteed to be complete [16].

To achieve a better understanding of the mimicry mechanism that possibly explains virus-infectious procedure, similarity statistical analysis can be performed according to z-score [66] and E-value [21] levels. Since the mimicry action occurs between host protein and pathogen protein, similarity statistics might help elucidate potential activities.

Overall, SIN, combined with binary protein-protein interactions, has many advantages for precise analysis based on statistics associated with 3D structure and domain information.

6 CHALLENGES

While the boom of big data analytics appears promising, when dealing with both the structural information and domain-domain interactions, there remains several challenges in the areas of SIN and HPPPI network development.

6.1 Feasible and Efficient Feature Representation

For computational models, especially protein sequences, feature representation remains a challenging topic. Various methods for feature representation currently exist [13], [14], [15], [20], [27], [28], [29], [30]. Previous results indicate that various representational methods yielded different performances across several species, although additional protein sequence information is being experimentally generated. We might observe this from the aspect of a small dataset (i.e. Clostridium botulinum and the big dataset: Bacillus anthracis).

Additional models based on deep learning techniques present end-to-end frameworks for learning from big data sets. The automatic feature extraction process could be a promising option for protein sequence research. Previously, we successfully employed a stacked denoising autoencoder as an unsupervised learning model to extract high-level feature for model learning [13]. Our result showed a potential direction for introducing deep learning neural networks.

Prior to inputting data into learning models, several traditional feature representation methods, including one-hot vector method, PSSM feature, and other statistic methods shown in TABLE 1, were widely used. Additionally, deep learning techniques are also first introduced in protein secondary structure prediction [22], [23] and HPPPI prediction tasks [13]. In terms of feature representation, deep learning techniques could harness the power of high-dimensional data in large volumes, enabling acquisition of large volumes of feature information to further improve model performance.

6.2 Imbalanced Data

Another challenging issue is the imbalanced ratio among different classes of the structural information, such as the eight categories of protein secondary structure. For structure prediction, domain-domain interaction and host-pathogen protein-protein interaction problems, the imbalanced ratio between different classes is important in improving model performance.

The ratio of non-interface interactions to interface interactions is about 9:1 [55]. In structure prediction task, the ratios in both Q3 and Q8 tasks are also different and imbalanced between different protein families. Specifically, for Q8 tasks, some structures are barely observable in the protein structures. In a previous study, the interacting pairs and non-interacting pairs were defined with 1:100 ratio, which is a highly skewed number [14].

With the continuous expansion and availability of structural information and domain data, the issues involving imbalanced data biological areas intensifies.

7 CONCLUSIONS

In this study, we presented a survey describing the building of structural interaction network (SIN) for host-pathogen protein-protein interactions to analyse the resulting network using a systems biology approach. We focused on structural information and also SIN analysis. Several multidisciplinary and interdisciplinary areas were reviewed, including protein feature representation, protein structure prediction, domain-domain interaction prediction and machine learning methods applied for these prediction tasks.

For HPPPI researches, building SIN using atomic level data can provide insights into high-resolution interactions

based on protein structures and offer high-quality analyses of interactions targeting infectious mechanisms. To the best of our knowledge, multiple areas still need to be addressed in this research direction. We anticipate this survey will benefit future proteomics studies, as well as the computational method design.

ACKNOWLEDGMENTS

This work is supported by the scholarship from the China Scholarship Council (CSC) and Faculty Strategic Investments Grant for DP 2019 development, while the first author pursues his PhD degree in the University of Wollongong.

REFERENCES

[1] H. Chen, J. Song, G. Sun, J. Shen, and L. Wang, "Towards elucidating the structural principles of host-pathogen protein-protein interaction networks: A bioinformatics survey," in *Big Data (BigData Congress), 2017 IEEE International Congress on*. IEEE, 2017, pp. 177–184.

[2] S. Orchard, H. Hermjakob, and R. Apweiler, "Proteomics and data standardisation," *Drug Discovery Today: Biosilico*, vol. 2, no. 3, pp. 91–93, 2004.

[3] U. Consortium *et al.*, "Uniprot: the universal protein knowledgebase," *Nucleic acids research*, vol. 45, no. D1, pp. D158–D169, 2017.

[4] M. Vaudel, K. Verheggen, A. Csordas, H. Ræder, F. S. Berven, L. Martens, J. A. Vizcaíno, and H. Barsnes, "Exploring the potential of public proteomics data," *Proteomics*, vol. 16, no. 2, pp. 214–225, 2016.

[5] A. Alyass, D. Meyre, and M. Turcotte, "From big data analysis to personalized medicine for all: challenges and opportunities," *BMC medical genomics*, vol. 8, no. 1, p. 33, 2015.

[6] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in bioinformatics*, vol. 18, no. 5, pp. 851–869, 2017.

[7] R. Breitling, "What is systems biology?" *Frontiers in physiology*, vol. 1, 2010.

[8] P. Aloy and R. B. Russell, "Structural systems biology: modelling protein interactions," *Nature Reviews Molecular Cell Biology*, vol. 7, no. 3, pp. 188–197, 2006.

[9] R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat, G. M. Mastrogianakis, J. J. Olson, T. Mikkelsen, N. Lehman, K. Aldape *et al.*, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, no. 7216, pp. 1061–1068, 2008.

[10] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Reviews Genetics*, vol. 16, no. 6, pp. 321–332, 2015.

[11] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, "Predicting protein–protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences*, vol. 104, no. 11, pp. 4337–4341, 2007.

[12] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences," *Nucleic acids research*, vol. 36, no. 9, pp. 3025–3030, 2008.

[13] H. Chen, J. Shen, L. Wang, and J. Song, "Leveraging stacked denoising autoencoder in prediction of pathogen-host protein-protein interactions," in *Big Data (BigData Congress), 2017 IEEE International Congress on*. IEEE, 2017, pp. 368–375.

[14] ——, "Towards data analytics of pathogen-host protein-protein interaction: a survey," in *Big Data (BigData Congress), 2016 IEEE International Congress on*. IEEE, 2016, pp. 377–388.

[15] ——, "Collaborative data analytics towards prediction on pathogen-host protein-protein interactions," in *International Conference on Computer Supported Cooperative Work in Design*. IEEE, 2017, pp. 269–274.

[16] E. A. Franzosa and Y. Xia, "Structural principles within the human-virus protein-protein interaction network," *Proceedings of the National Academy of Sciences*, vol. 108, no. 26, pp. 10 538–10 543, 2011.

[17] K. Li, C. Xu, J. Huang, W. Liu, L. Zhang, W. Wan, H. Tao, L. Li, S. Lin, A. Harrison *et al.*, "Prediction and identification of the effectors of heterotrimeric g proteins in rice (oryza sativa l.)," *Briefings in bioinformatics*, vol. 18, no. 2, pp. 270–278, 2016.

[18] M. N. Davies, A. Secker, A. A. Freitas, E. Clark, J. Timmis, and D. R. Flower, "Optimizing amino acid groupings for gpcr classification," *Bioinformatics*, vol. 24, no. 18, pp. 1980–1986, 2008.

[19] X. Lin, X.-w. Chen *et al.*, "On position-specific scoring matrix for protein function prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 8, no. 2, pp. 308–315, 2011.

[20] M. Bhagwat and L. Aravind, "Psi-blast tutorial," *Comparative Genomics*, pp. 177–186, 2008.

[21] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[22] J. Zhou and O. Troyanskaya, "Deep supervised and convolutional generative stochastic network for protein secondary structure prediction," in *International Conference on Machine Learning*, 2014, pp. 745–753.

[23] Z. Li and Y. Yu, "Protein secondary structure prediction using cascaded convolutional and recurrent neural networks," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 2560–2567.

[24] M. Zamani and S. C. Kremer, "Protein secondary structure prediction using an evolutionary computation method and clustering," in *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on*. IEEE, 2015, pp. 1–6.

[25] C. Floudas, H. Fung, S. McAllister, M. Mönnigmann, and R. Rajgaria, "Advances in protein structure prediction and de novo protein design: A review," *Chemical Engineering Science*, vol. 61, no. 3, pp. 966–988, 2006.

[26] Z. Aydin, Y. Altunbasak, and M. Borodovsky, "Protein secondary structure prediction for a single-sequence using hidden semi-markov models." *BMC bioinformatics*, vol. 7, pp. 178–178, 2006.

[27] M. Spencer, J. Eickholt, and J. Cheng, "A deep learning network approach to ab initio protein secondary structure prediction," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 12, no. 1, pp. 103–112, 2015.

[28] A. Yaseen and Y. Li, "Template-based c8-scorpion: a protein 8-state secondary structure prediction method using structural information and context-based features," *BMC bioinformatics*, vol. 15, no. 8, p. S3, 2014.

[29] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *Journal of molecular biology*, vol. 202, no. 4, pp. 865–884, 1988.

[30] O. Dor and Y. Zhou, "Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training," *Proteins: Structure, Function, and Bioinformatics*, vol. 66, no. 4, pp. 838–845, 2007.

[31] B. Jayaram, P. Dhingra, A. Mishra, R. Kaushik, G. Mukherjee, A. Singh, and S. Shekhar, "Bhageerath-h: A homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins," *BMC bioinformatics*, vol. 15, no. S16, p. S7, 2014.

[32] S. Kaczanowski and P. Zielenkiewicz, "Why similar protein sequences encode similar three-dimensional structures?" *Theoretical Chemistry Accounts*, vol. 125, no. 3-6, pp. 643–650, 2010.

[33] S. Yellaboina, A. Tasneem, D. V. Zaykin, B. Raghavachari, and R. Jothi, "Domine: a comprehensive collection of known and predicted domain-domain interactions," *Nucleic acids research*, vol. 39, no. suppl_1, pp. D730–D735, 2010.

[34] A. Stein, A. Panjkovich, and P. Aloy, "3did update: domain–domain and peptide-mediated interactions of known 3d structure," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D300–D304, 2008.

[35] R. D. Finn, B. L. Miller, J. Clements, and A. Bateman, "ipfam: a database of protein family and domain interactions found in the protein data bank," *Nucleic acids research*, vol. 42, no. D1, pp. D364–D373, 2013.

[36] B. A. Shoemaker and A. R. Panchenko, "Deciphering protein–protein interactions. part ii. computational methods to predict protein and domain interaction partners," *PLoS computational biology*, vol. 3, no. 4, p. e43, 2007.

[37] S. Khor, "Inferring domain-domain interactions from protein-protein interactions with formal concept analysis," *PloS one*, vol. 9, no. 2, p. e88943, 2014.

[38] X.-M. Zhao, G. Chesi, and L. Chen, "Computational systems biology: Understanding biological systems from the perspective of networks and dynamics," *IEEE Systems, Man and Cybernetics Society: eNewsletter*, vol. 26, March 2009.

[39] J. L. Sussman, D. Lin, J. Jiang, N. O. Manning, J. Prilusky, O. Ritter, and E. Abola, "Protein data bank (pdb): database of three-dimensional structural information of biological macromolecules," *Acta Crystallographica Section D: Biological Crystallography*, vol. 54, no. 6, pp. 1078–1084, 1998.

[40] T. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal *et al.*, "Human protein reference database2009 update," *Nucleic acids research*, vol. 37, no. suppl_1, pp. D767–D772, 2008.

[41] G. D. Bader, D. Betel, and C. W. Hogue, "Bind: the biomolecular interaction network database," *Nucleic acids research*, vol. 31, no. 1, pp. 248–250, 2003.

[42] I. Xenarios, L. Salwinski, X. J. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions," *Nucleic acids research*, vol. 30, no. 1, pp. 303–305, 2002.

[43] A. R. Wattam, D. Abraham, O. Dalay, T. L. Disz, T. Driscoll, J. L. Gabbard, J. J. Gillespie, R. Gough, D. Hix, R. Kenyon *et al.*, "Patric, the bacterial bioinformatics database and analysis resource," *Nucleic acids research*, vol. 42, no. D1, pp. D581–D591, 2013.

[44] S. Durmuş Tekir, T. Çakır, E. Ardıç, A. S. Sayılırbaş, G. Konuk, M. Konuk, H. Sarıyer, A. Uğurlu, İ. Karadeniz, A. Özgür *et al.*, "Phisto: pathogen–host interaction search tool," *Bioinformatics*, vol. 29, no. 10, pp. 1357–1358, 2013.

[45] W. L. DeLano, "The pymol molecular graphics system," *http://pymol.org*, 2002.

[46] Schrödinger, LLC, "The PyMOL molecular graphics system, version 1.8," *https://pymol.org/2/*, November 2015.

[47] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry *et al.*, "Pfam: the protein families database," *Nucleic acids research*, vol. 42, no. D1, pp. D222–D230, 2013.

[48] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas *et al.*, "The pfam protein families database: towards a more sustainable future," *Nucleic acids research*, vol. 44, no. D1, pp. D279–D285, 2016.

[49] R. Mosca, A. Céol, A. Stein, R. Olivella, and P. Aloy, "3did: a catalog of domain-based interactions of known three-dimensional structure," *Nucleic acids research*, vol. 42, no. D1, pp. D374–D379, 2013.

[50] P. Stolorz, A. Lapedes, and Y. Xia, "Predicting protein secondary structure using neural net and statistical methods," *Journal of Molecular Biology*, vol. 225, no. 2, pp. 363–377, 1992.

[51] J. Garnier, J.-F. Gibrat, and B. Robson, "[32] gor method for predicting protein secondary structure from amino acid sequence," *Methods in enzymology*, vol. 266, pp. 540–553, 1996.

[52] T. Z. Sen, R. L. Jernigan, J. Garnier, and A. Kloczkowski, "Gor v server for protein secondary structure prediction," *Bioinformatics*, vol. 21, no. 11, pp. 2787–2788, 2005.

[53] S. Hua and Z. Sun, "A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach," *Journal of molecular biology*, vol. 308, no. 2, pp. 397–407, 2001.

[54] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[55] X.-w. Chen and J. C. Jeong, "Sequence-based prediction of protein interaction sites with an integrative method," *Bioinformatics*, vol. 25, no. 5, pp. 585–591, 2009.

[56] Z.-S. Wei, K. Han, J.-Y. Yang, H.-B. Shen, and D.-J. Yu, "Protein–protein interaction sites prediction by ensembling svm and sample-weighted random forests," *Neurocomputing*, vol. 193, pp. 201–212, 2016.

[57] H. Hu, J. Li, H. Wang, G. Daggard, and M. Shi, "A maximally diversified multiple decision tree algorithm for microarray data classification," in *Proceedings of the 2006 workshop on Intelligent systems for bioinformatics-Volume 73*. Australian Computer Society, Inc., 2006, pp. 35–38.

[58] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[59] B. Rost, "[31] phd: Predicting one-dimensional protein structure by profile-based neural networks," *Methods in enzymology*, vol. 266, pp. 525–539, 1996.

[60] Z. Wang, F. Zhao, J. Peng, and J. Xu, "Protein 8-class secondary structure prediction using conditional neural fields," *Proteomics*, vol. 11, no. 19, pp. 3786–3792, 2011.

[61] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, and Y. Zhou, "Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning," *Scientific reports*, vol. 5, p. 11476, 2015.

[62] X. Wang, X. Wei, B. Thijssen, J. Das, S. M. Lipkin, and H. Yu, "Three-dimensional reconstruction of protein networks provides insight into human genetic disease," *Nature biotechnology*, vol. 30, no. 2, pp. 159–164, 2012.

[63] P. D. Stenson, M. Mort, E. V. Ball, K. Howells, A. D. Phillips, N. S. Thomas, and D. N. Cooper, "The human gene mutation database: 2008 update," *Genome medicine*, vol. 1, no. 1, p. 13, 2009.

[64] E. A. Franzosa, S. Garamszegi, and Y. Xia, "Toward a three-dimensional view of protein networks between species," *Frontiers in microbiology*, vol. 3, 2012.

[65] E. Guven-Maiorov, C.-J. Tsai, and R. Nussinov, "Structural host-microbiota interaction networks," *PLoS computational biology*, vol. 13, no. 10, p. e1005579, 2017.

[66] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," *Journal of molecular biology*, vol. 233, no. 1, pp. 123–138, 1993.